Supplementary Document: Proofs

Shahira Amin, Yuanxiong Guo, and Yanmin Gong

Appendix A
Proof of Theorem III.1:

We denote $[k]$ as the time interval between two successive aggregations $t \in [(k-1)\tau, k\tau), k = 1, \dots, K$. We define a *centralized* model that follows a centralized gradient descent update as

$$\mathbf{v}_{[k]}(t) = \mathbf{v}_{[k]}(t-1) - \eta \nabla L(\mathbf{v}_{[k]}(t-1)), \quad \forall t \in [(k-1)\tau, k\tau). \tag{1}$$

This update is performed at base station, where all data samples collected at edge devices D_W, the global machine learning loss function, and its gradient are assumed to be available.

We assume that the local loss $L_i(\mathbf{w}), \forall i \in \mathcal{I}$ is convex, ρ-Lipschitz, and β-smooth. Furthermore, we let the following conditions hold: 1) $\eta \leq \frac{1}{\beta}$, 2) $L(\mathbf{v}_{[k]}(k\tau)) - L(\mathbf{w}^*) \geq \epsilon$, and 3) $L(\mathbf{w}_i(t)) - L(\mathbf{w}^*) \geq \epsilon$. for $\epsilon > 0$.

At the beginning of interval $[k]$, we synchronize the centralized model $\mathbf{v}_{[k]}((k-1)\tau)$ with the global model $\mathbf{w}((k-1)\tau)$ in (9), i.e., $\mathbf{v}_{[k]}((k-1)\tau) = \mathbf{w}((k-1)\tau)$. We define $\theta_{[k]}(t) = L(\mathbf{v}_{[k]}(t)) - L(\mathbf{w}^*)$, where $\mathbf{w}^*$ is the optimal solution. Considering $K = \frac{T}{\tau}$:

$$\begin{aligned}
\frac{1}{\theta_{[K+1]}(t)} - \frac{1}{\theta_{[1]}(0)} &= \sum_{y=K\tau+1}^{t} \left(\frac{1}{\theta_{[K+1]}(y)} - \frac{1}{\theta_{[K+1]}(y-1)} \right) + \left(\frac{1}{\theta_{[K+1]}(K\tau)} - \frac{1}{\theta_{[K]}(K\tau)} \right) + \left(\frac{1}{\theta_{[K]}(K\tau)} - \frac{1}{\theta_{[1]}(0)} \right) \\
&\geq \Big((t - K\tau)\omega\eta\Big(1 - \frac{\beta\eta}{2}\Big)\Big) + \Big(\frac{-\rho h(\tau)}{\epsilon^2} \Big) + \Big(T\omega\eta\Big(1 - \frac{\beta\eta}{2}\Big) - (K-1)\frac{\rho h(\tau)}{\epsilon^2} \Big) \quad \text{(Lemma 2 in [1])} \\
&= t\omega\eta\Big(1 - \frac{\beta\eta}{2}\Big) - K\frac{\rho h(\tau)}{\epsilon^2},
\end{aligned} \tag{2}$$

where $\omega = \min_k \frac{1}{\|\mathbf{v}_{[k]}((k-1)\tau) - \mathbf{w}^*\|^2}$ and $h(x) \triangleq \frac{\delta}{\beta}((\eta\beta + 1)^x - 1) - \eta\delta x, \forall x \in \{0, 1, \dots\}$.

Letting $\theta_{[k]}(k\tau) = L(\mathbf{v}_{[k]}(k\tau)) - L(\mathbf{w}^*) \geq \epsilon, \forall k$. From Lemma 5 in [1], $L(\mathbf{v}_{[k]}(t)) \geq L(\mathbf{v}_{[k]}(t+1)), \forall t \in [(k-1)\tau, k\tau)$. Hence, $\theta_{[k]}(t) = L(\mathbf{v}_{[k]}(t)) - L(\mathbf{w}^*) \geq \epsilon, \forall t, k$ at which $\mathbf{v}_{[k]}(t)$ is defined. Additionally, we assume that $L(\mathbf{w}_i(t)) - L(\mathbf{w}^*) \geq \epsilon$. Therefore,

$$\begin{aligned}
\frac{1}{L(\mathbf{w}_i(t)) - L(\mathbf{w}^*)} - \frac{1}{\theta_{[K+1]}(t)} &= \frac{\theta_{[K+1]}(t) - (L(\mathbf{w}_i(t)) - L(\mathbf{w}^*))}{(L(\mathbf{w}_i(t)) - L(\mathbf{w}^*))\theta_{[K+1]}(t)} \\
&= \frac{L(\mathbf{v}_{[K+1]}(t)) - L(\mathbf{w}_i(t))}{(L(\mathbf{w}_i(t)) - L(\mathbf{w}^*))\theta_{[K+1]}(t)} \geq -\frac{\rho g_i(t - K\tau)}{\epsilon^2} \\
&\qquad \text{(from Lemma 3 in [1] and Lipschitz condition)},
\end{aligned} \tag{3}$$

where $g_i(x) \triangleq \frac{\delta_i}{\beta}((\eta\beta + 1)^x - 1)$. By adding (2) and (3), and assuming $\theta_{[k]}(t) \geq 0$ (according to Theorem 3.14 in [2]):

$$\frac{1}{L(\mathbf{w}_i(t)) - L(\mathbf{w}^*)} \geq \frac{1}{L(\mathbf{w}_i(t)) - L(\mathbf{w}^*)} - \frac{1}{\theta_{[1]}(0)} \geq t\omega\eta\Big(1 - \frac{\beta\eta}{2}\Big) - \frac{\rho}{\epsilon^2}(Kh(\tau) + g_i(t - K\tau)). \tag{4}$$

Consequently,

$$L(\mathbf{w}_i(t)) - L(\mathbf{w}^*) \leq \frac{1}{t\omega\eta\Big(1 - \frac{\beta\eta}{2}\Big) - \frac{\rho}{\epsilon^2}(Kh(\tau) + g_i(t - K\tau))} = y(\epsilon). \tag{5}$$

Solving $y(\epsilon_0) = \epsilon_0$, we obtain the positive solution of ϵ_0 as

$$\epsilon_0 = \frac{1}{t\omega\eta(2 - \beta\eta)} + \sqrt{\frac{1}{t^2\omega^2\eta^2(2 - \beta\eta)^2} + \frac{Kh(\tau) + g_i(t - K\tau)}{t\omega\eta(1 - \frac{\beta\eta}{2})}}. \tag{6}$$

We ignore the negative solution of ϵ_0 because $L(\mathbf{w}_i(t)) - L(\mathbf{w}^*) \geq \epsilon > 0$. Assuming that there exists $\epsilon > \epsilon_0$ satisfying:

$$L(\mathbf{v}_{[k]}(k\tau)) - L(\mathbf{w}^*) \geq \epsilon \tag{7}$$

$$L(\mathbf{w}_i(t)) - L(\mathbf{w}^*) \geq \epsilon \tag{8}$$

Then,

$$L(\mathbf{w}_i(t)) - L(\mathbf{w}^*) \le \frac{1}{t\omega\eta\Big(1-\frac{\beta\eta}{2}\Big) - \frac{\rho}{\epsilon^2}(Kh(\tau)+g_i(t-K\tau))}$$
$$\le \frac{1}{t\omega\eta\Big(1-\frac{\beta\eta}{2}\Big) - \frac{\rho}{\epsilon_0^2}(Kh(\tau)+g_i(t-K\tau))} = \epsilon_0 < \epsilon. \tag{9}$$

This is because the denominator in (9) is increasing with ϵ when $\rho(Kh(\tau)+g_i(t-K\tau)) > 0$. Due to the contradiction between (8) and (9), we can conclude that there *does not* exist $\epsilon > \epsilon_0$ satisfying both (7) and (8). Therefore, one of the following conditions must hold: either 1) $L(\mathbf{v}_{[k]}(k\tau)) - L(\mathbf{w}^*) \le \epsilon_0$ or 2) $L(\mathbf{w}_i(t)) - L(\mathbf{w}^*) \le \epsilon_0$.

Let the first condition holds. We could re-write $L(\mathbf{v}_{[k]}(k\tau)) - L(\mathbf{w}^*) \le \epsilon_0$ as follows: $\min_k L(\mathbf{v}_{[k]}(k\tau)) - L(\mathbf{w}^*) \le \epsilon_0$. Since $L(\mathbf{v}_{[k]}(k\tau))$ is non-increasing with k, therefore, $L(\mathbf{v}_{[K+1]}(K\tau)) - L(\mathbf{w}^*) \le \epsilon_0$. From Lemma 3 in [1], $\|\mathbf{w}_i(t) - \mathbf{v}_{[k]}(t)\| \le g_i(t-(k-1)\tau)$. Hence, using Lipschitz condition, $L(\mathbf{w}_i(t)) - L(\mathbf{v}_{[K+1]}(t)) \le \rho g_i(t-K\tau)$. Thus, $L(\mathbf{w}_i(t)) - L(\mathbf{w}^*) \le L(\mathbf{v}_{[K+1]}(t)) - L(\mathbf{w}^*) + \rho g_i(t-K\tau)$. Since the first condition holds, then $L(\mathbf{w}_i(t)) - L(\mathbf{w}^*) \le \epsilon_0 + \rho g_i(t-K\tau)$.

Let the second condition holds. Therefore, $L(\mathbf{w}_i(t)) - L(\mathbf{w}^*) \le \epsilon_0 + \rho g_i(t-K\tau)$, since $\rho > 0$ and $g_i(t-K\tau) > 0$. Hence, we can conclude that either the first condition or the second condition implies that:

$$L(\mathbf{w}_i(t)) - L(\mathbf{w}^*) \le \epsilon_0 + \rho g_i(t-K\tau). \tag{10}$$

Using the triangle inequality,

$$\|\nabla L_i(\mathbf{w}|\mathcal{G}_i(t)) - \nabla L(\mathbf{w})\| = \|\nabla L_i(\mathbf{w}|\mathcal{G}_i(t)) - \nabla L(\mathbf{w}|\mathcal{D}_i(t)) + \nabla L(\mathbf{w}|\mathcal{D}_i(t)) - \nabla L(\mathbf{w}|\mathcal{D}) + \nabla L(\mathbf{w}|\mathcal{D}) - \nabla L(\mathbf{w}|\mathcal{D}_W)\|$$
$$\le \|\nabla L_i(\mathbf{w}|\mathcal{G}_i(t)) - \nabla L(\mathbf{w}|\mathcal{D}_i(t))\| + e + \|\nabla L(\mathbf{w}|\mathcal{D}) - \nabla L(\mathbf{w}|\mathcal{D}_W)\|. \tag{11}$$

Using the central limit theorem, since $\nabla L(\mathbf{w}|\mathcal{D}_W)$ is the sample average of $\nabla L(\mathbf{w}, \mathbf{x}_d, y_d), \forall(\mathbf{x}_d, y_d) \in \mathcal{D}_W$, then $\nabla L(\mathbf{w}|\mathcal{D}_W)$ can be regarded as D_W samples drawn from a distribution whose mean value is $\nabla L(\mathbf{w}|\mathcal{D})$. Therefore, $\nabla L(\mathbf{w}|\mathcal{D}_i(t)) - \nabla L(\mathbf{w}|\mathcal{D})$ could be upper bounded as

$$\|\nabla L(\mathbf{w}|D) - \nabla L(\mathbf{w}|D_W)\| \le \frac{\gamma}{\sqrt{D_W}} \tag{12}$$

Similarly, by applying the central limit theorem, $\|\nabla L_i(\mathbf{w}|\mathcal{G}_i(t)) - \nabla L(\mathbf{w}|\mathcal{D}_i(t))\|$ could be upper bounded as

$$\|\nabla L_i(\mathbf{w}|\mathcal{G}_i(t)) - \nabla L(\mathbf{w}|\mathcal{D}_i(t))\| \le \frac{\gamma_i}{\sqrt{G_i(t)}}, \tag{13}$$

where $\gamma_i > 0$ is a constant that does not depend on $G_i(t)$. Substituting (12) and (13) in (11),

$$\|\nabla L_i(\mathbf{w}|\mathcal{G}_i(t)) - \nabla L(\mathbf{w})\| \le \frac{\gamma_i}{\sqrt{G_i(t)}} + \frac{\gamma}{\sqrt{D_W}} + e \tag{14}$$

Defining a constant δ_i as an upper bound for the gradient divergence $\|\nabla L_i(\mathbf{w}) - \nabla L(\mathbf{w})\| \le \delta_i$, such that $\delta = \frac{\sum_{i\in\mathcal{I}} D_i\delta_i}{D_W}$ ($D_i = \cup_{t\le T} D_i(t)$). From (10), $L(\mathbf{w}_i(t)) - L(\mathbf{w}^*(t)) \propto g_i(x) \propto \delta_i$, and from (14), $\delta_i \equiv \|\nabla L_i(\mathbf{w}|\mathcal{G}_i(t)) - \nabla L(\mathbf{w})\| \propto \sqrt{G_i^{-1}(t)}$. Therefore,

$$L(\mathbf{w}_i(t)) - L(\mathbf{w}^*(t)) \propto \sqrt{G_i^{-1}(t)} \tag{15}$$

APPENDIX B
PROOF OF LEMMA IV.1

To get an upper for the *drift-plus-penalty*, we use the inequality: $((Y-b)^+ + A)^2 \le Y^2 + A^2 + b^2 + 2Y(A-b)$. Applying this inequality to (3), we obtain:

$$Q_i^2(t+1) \le Q_i^2(t) + \Big(\sum_{j=1}^{I} f_{ji}(t)\Big)^2 + 2Q_i(t)\Big(\sum_{j=1}^{I} f_{ji}(t) - G_i(t)\Big) + G_i(t)^2 \tag{16}$$

Taking the sum over all devices:

$$\sum_{i=1}^{I}\frac{Q_i^2(t+1)}{2} - \sum_{i=1}^{I}\frac{Q_i^2(t)}{2} \le \sum_{i=1}^{I}\frac{\Big(\sum_{j=1}^{I} f_{ji}(t)\Big)^2 + G_i(t)^2}{2} + \sum_{i=1}^{I} Q_i(t)\Big(\sum_{j=1}^{I} f_{ji}(t) - G_i(t)\Big) \tag{17}$$

Taking the conditional expectation of (17):

$$\Delta(\mathbf{Q}(t)) \le B_1 + \sum_{i=1}^{I} \mathbb{E}\Big\{Q_i(t)\Big(\sum_{j=1}^{I} f_{ji}(t) - G_i(t)\Big)|\mathbf{Q}(t)\Big\}, \tag{18}$$

where B_1 is a constant $B_1 = \frac{1}{2}\sum_{i=1}^{I}\big(\sum_{j=1}^{I} B_{ij}\big)^2 + C_i^2$. Summing $V\mathbb{E}\{Cost(t)|\mathbf{Q}(t)\}$ to both sides of (18), we obtain an upper bound for the *drift-plus-penalty*:

$$\Delta_v(t) \triangleq \Delta(\mathbf{Q}(t)) + V\mathbb{E}\{Cost(t)|\mathbf{Q}(t)\} \le B_1 + V\mathbb{E}\{Cost(t)|\mathbf{Q}(t)\} + \sum_{i=1}^{I} \mathbb{E}\Big\{Q_i(t)\Big(\sum_{j=1}^{I} f_{ji}(t) - G_i(t)\Big)|\mathbf{Q}(t)\Big\} \tag{19}$$

Appendix C
Proof of Lemma IV.2

Let $f_{ij}^*(t)$ and $G_i^*(t)$ be the optimal solution of problem (15). Therefore,

$$\sum_{i=1}^{I}\Big[Q_i(t)\Big(\sum_{j=1}^{I} f_{ji}^*(t) - G_i^*(t)\Big)\Big] + VCost(t) \le \sum_{i=1}^{I}\Big[Q_i(t)\Big(\sum_{j=1}^{I} f_{ji}(t) - G_i(t)\Big)\Big] + VCost(t) \tag{20}$$

Taking the conditional expectation of (20):

$$\begin{aligned}\sum_{i=1}^{I}\Big[\mathbb{E}\Big\{Q_i(t)\Big(\sum_{j=1}^{I} f_{ji}^*(t) - G_i^*(t)\Big)|\mathbf{Q}(t)\Big\}\Big] + V\mathbb{E}\{Cost(t)|\mathbf{Q}(t)\}&\\ \le \sum_{i=1}^{I}\Big[\mathbb{E}\Big\{Q_i(t)\Big(\sum_{j=1}^{I} f_{ji}(t) - G_i\Big)|\mathbf{Q}(t)\Big\}\Big] + V\mathbb{E}\{Cost(t)|\mathbf{Q}(t)\}.&\end{aligned} \tag{21}$$

Therefore, we can conclude that the optimal solution of problem (15) minimizes the upper bound of the *drift-plus-penalty*.

Appendix D
Proof of Theorem IV.3

If $(U_i(t), D_i(t), \forall i)$ is i.i.d. over time, and if there exists a constant ζ such that $\omega + \zeta\mathbf{1} \in \Omega$, where $\omega = \mathbb{E}\{D_i(t)\}$, then it can be shown that there exists a stationary and randomized policy $(f_{ij}^{'}(t), G_i^{'}(t), \forall i, j \in \mathcal{I})$ that satisfies the following:

$$\mathbb{E}\Big\{\sum_{j=1}^{I} f_{ji}^{'}(t)\Big\} = \mathbb{E}\Big\{G_i^{'}(t)\Big\} - \zeta, \forall i \tag{22}$$

$$\mathbb{E}\{Cost^{'}(t)\} = \overline{g}^{'}(\omega + \zeta\mathbf{1}), \tag{23}$$

where $Cost^{'}(t)$ is the cost function evaluated at $f_{ij}^{'}(t), G_i^{'}(t), \forall i, j \in \mathcal{I}$ and $\overline{g}^{'}$ is the optimal cost.Therefore, from (19), we have:

$$\begin{aligned}\Delta_v(t) \triangleq \Delta(\mathbf{Q}(t)) + V\mathbb{E}\{Cost(t)|\mathbf{Q}(t)\} &\le B_1 + V\mathbb{E}\{Cost^{'}(t)|\mathbf{Q}(t)\} + \sum_{i=1}^{I}\mathbb{E}\Big\{Q_i(t)\Big(\sum_{j=1}^{I} f_{ji}^{'}(t) - G_i^{'}(t)\Big)|\mathbf{Q}(t)\Big\}\\ &\le B_1 + V\sum_{i=1}^{I}\mathbb{E}\{Cost^{'}(t)\} + \sum_{i=1}^{I} Q_i(t)\mathbb{E}\Big\{\sum_{j=1}^{I} f_{ji}^{'}(t) - G_i^{'}(t)\Big\},\end{aligned} \tag{24}$$

where we have used the assumption that $(U_i(t), D_i(t), \forall i)$ is i.i.d. over time. Therefore, $Cost^{'}(t)$, $f_{ji}^{'}(t)$, and $G_i^{'}(t)$ are independent of queue backlog. From (22) and (23):

$$\Delta_v(t) \le B_1 + V\overline{g}^{'}(\omega + \zeta\mathbf{1}) - \zeta\sum_{i=1}^{I} Q_i(t) \tag{25}$$

Taking the expectation of both sides

$$\mathbb{E}\{\mathbf{Y}(t+1) - \mathbf{Y}(t)\} + V\mathbb{E}\{Cost(t)\} \le B_1 + V\overline{g}^{'}(\omega + \zeta\mathbf{1}) - \zeta\sum_{i=1}^{I}\mathbb{E}\{Q_i(t)\} \tag{26}$$

Using the law of iterative expectation and the fact that $\overline{g}^{'}(\omega + \delta\mathbf{1}) \leq B_2$, where B_2 is the upper bound for the cost function and could be expressed as:

$$Cost(t) \leq B_2 = \sum_{i=1}^{I} \Big(U_i^{max} \Big(\sum_{j=1}^{I} c_{ij} B_{ij} + c_i C_i \Big) + L_i^{max} \Big) \tag{27}$$

Then, taking the time-average of (26),

$$\frac{1}{T}\mathbb{E}\big\{\mathbf{Y}(T)\big\} - \frac{1}{T}\mathbb{E}\big\{\mathbf{Y}(0)\big\} + \frac{V}{T}\mathbb{E}\big\{Cost(t)\big\} \leq B_1 + VB_2 - \frac{\zeta}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big\{\sum_{i=1}^{I} Q_i(t)\Big\} \tag{28}$$

Since $\mathbf{Y}(0) = 0$ because the queues are initially empty and $\mathbf{Y}(T) \geq 0$, therefore, taking the limit as $T \to \infty$, we get:

$$\overline{Q} \leq \frac{B_1 + VB_2}{\zeta} \tag{29}$$

To prove the bound on the cost performance, from (26), we have:

$$\mathbb{E}\big\{\mathbf{Y}(t+1) - \mathbf{Y}(t)\big\} + V\mathbb{E}\big\{Cost(t)\big\} \leq B_1 + V\overline{g}^{'}(\omega + \zeta\mathbf{1}) \tag{30}$$

Taking the time-average, and using the fact that $\mathbf{Y}(0) = 0$ and $\mathbf{Y}(T) \geq 0$, we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big\{Cost(t)\big\} \leq \frac{B_1}{V} + \overline{g}^{'}(\omega + \zeta\mathbf{1}) \tag{31}$$

Letting $T \to \infty$, we obtain an upper bound on time-average cost as:

$$\overline{Cost}(t) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big\{Cost(t)\big\} \leq \frac{B_1}{V} + \overline{g}^{*} \tag{32}$$

where $\overline{g}^{*}$ is the optimal objective value achieved by any control policy.

References

[1] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[2] S. Bubeck, "Convex optimization: Algorithms and complexity," *arXiv preprint arXiv:1405.4980*, 2014.